CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models

Year: 2020 | Citations: 328 | Authors: Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao

Abstract

Learning disentanglement aims at finding a low dimensional representation which consists of multiple explanatory and generative factors of the observational data. The framework of variational autoencoder (VAE) is commonly used to disentangle independent factors from observations. However, in real scenarios, factors with semantics are not necessarily independent. Instead, there might be an underlying causal structure which renders these factors dependent. We thus propose a new VAE based framework named CausalVAE, which includes a Causal Layer to transform independent exogenous factors into causal endogenous ones that correspond to causally related concepts in data. We further analyze the model identifiabitily, showing that the proposed model learned from observations recovers the true one up to a certain degree. Experiments are conducted on various datasets, including synthetic and real word benchmark CelebA. Results show that the causal representations learned by CausalVAE are semantically interpretable, and their causal relationship as a Directed Acyclic Graph (DAG) is identified with good accuracy. Furthermore, we demonstrate that the proposed CausalVAE model is able to generate counterfactual data through "do-operation" to the causal factors.